Metrika (2011) 74:67–83
DOI 10.1007/s00184-009-0290-z

Control charts for health care monitoring under overdispersion

Willem Albers

Received: 21 April 2009 / Published online: 29 October 2009

Abstract An attractive way to control attribute data from high quality processes is to wait till $r \geq 1$ failures have occurred. The choice of r in such negative binomial charts is dictated by how much the failure rate is supposed to change during Out-of-Control. However, these results have been derived for the case of homogeneous data. Especially in health care monitoring, (groups of) patients will often show large heterogeneity. In the present paper we will show how such overdispersion can be taken into account. In practice, typically neither the average failure rate, nor the overdispersion parameter(s), will be known. Hence we shall also derive and analyze the estimated version of the new chart.

Keywords Statistical process control · High-quality processes · Geometric charts · Average run length · Estimated parameters · Heterogeneity

Mathematics Subject Classification (2000) 62P10 · 62C05 · 62F12

1 Introduction and motivation

In this paper we consider high-quality processes, in which the proportion of non-conforming items is expected to be (very) small. First of all, due to constant efforts to improve quality in production, such a setup will be encountered more and more often in industrial settings. Moreover, in the quite different, but equally important, field of health care monitoring, this is in fact the standard situation: negative events

W. Albers (✉)
Department of Applied Mathematics, University of Twente,
P.O. Box 217, 7500 AE Enschede, The Netherlands
e-mail: w.albers@utwente.nl

[malfunctioning equipment, unsuccessful surgery, excessive delay before help arrives, detection of (the return of) a serious disease] should typically be (very) rare.

In review papers on health care monitoring [see e.g. Woodall (2006); Thor et al. (2007); Shaha (1995) and Sonesson and Bock (2003)], the use of SPC methods is strongly advocated, with special emphasis on control charts as the key tools. Now a standard choice for controlling attribute data is a p-chart, based on the number of failures in a series of given sampling intervals. However, for the really small proportions p we encounter in high-quality processes, substantial improvements can be achieved by applying a different type of chart, which goes by a variety of names, such as 'time-between-events' or 'geometric'. All such charts essentially employ the number of successes between failures, see e.g. Liu et al. (2004); Yang et al. (2002); Xie et al. (1998); Ohta et al. (2001); Zhang et al. (2004) and Wu et al. (2001).

A known drawback of this geometric chart, however, is that it requires a rather long time to react to a moderate increase of the failure rate p. Only large deteriorations quickly produce an Out-of-Control (OoC) signal. Clearly, in particular for health care applications, this can be quite unacceptable. Most of the authors quoted above [and also Bourke (1991, 2006)] therefore suggest as a remedy to essentially use a negative binomial chart: postpone the decision whether to stop until $r > 1$ failures have occurred. Some guidance on how to choose r in practice can be found in Ohta et al. (2001), but a systematic treatment of this issue was given in Albers (2010), resulting in a simple rule of thumb for choosing the optimal r as a function of the desired false alarm rate (FAR) and the supposed degree of increase of p compared to its value during In-Control (IC). As expected, the larger the increase one has in mind, the smaller r should be, with again the geometric chart ($r = 1$) as the ultimate result. In passing, we mention that yet another way to extend the geometric chart is the so-called sets method introduced by Chen (1978) [also see Gallus et al. (1986) and Chen (1987)]. Here the criterion is not the sum of the numbers of successes in r consecutive intervals between failures, but rather their maximum.

The second problem addressed in Albers (2010) concerns the estimation step involved. Note the general nature of this issue: typically, control charts have one or more unknown parameters which first have to be estimated on the basis of a so-called Phase I sample. Contrary to popular optimism, the effects of this estimation step are only negligible when (much) larger sample sizes are used than is customary in practice. Hence as a rule, such effects have to be taken into account and, if possible, corrections should be applied to the control limits to neutralize these. This program is indeed carried out in Albers (2010) for the negative binomial charts when p is unknown, and the result is a chart which is both simple to understand and to apply.

As such it thus offers a very satisfactory solution to the problem of monitoring high quality processes, characterized by an incoming sequence $D_1, D_2, \ldots$, of independent identically distributed (i.i.d.) random variables (r.v.'s) with $Pr(D_1 = 1) = 1 - Pr(D_1 = 0) = p$, where p is (very) small. However, note the underlying homogeneity assumption, which is made explicit by this more formal description. For industrial processes this assumption usually is quite reasonable, although it will certainly not always be warranted. But in medical applications, patients will often show large heterogeneity, and we really have to take such variation between subjects into account on a rather regular basis.

Roughly speaking two types of situations should be distinguished. In the first, we essentially only know that such heterogeneity does occur. It is e.g. due to the existence of different subgroups, each with its own probability of failure, but we lack further information. The only way in which it becomes apparent, is through an increase of variance over what would be expected under the homogeneous model. This is the well-known phenomenon of overdispersion. See e.g. Poortema (1999) for a general review, and more specifically in connection with attribute control charts, Christensen et al. (2003) and Fang (2003) for an industrial setting and Marshall et al. (2004) and Grigg et al. (2009) for health care monitoring applications. The present paper will be devoted to demonstrating how negative binomial charts can be adapted to cover the overdispersion situation as well.

However, before addressing this issue, in passing we consider the second of the two situations mentioned above. Here we do have knowledge about the underlying structure. For example, incoming patients are classified into different risk categories, for each of which the corresponding p_i is known or can be estimated. This opens the possibility for so-called risk adjustment [see Grigg and Farewell (2004a) for an overview and Grigg and Farewell (2004b) for a risk-adjusted version of the sets method]: the base-line risk of each patient can be taken into account, thus allowing a more accurate appraisal of e.g. a surgeon's performance on a series of such patients. Clearly, this is an interesting option, giving rise to various questions. For what type of application is risk adjustment advisable, how should it be applied, what are the (typically larger!) estimation effects and how can these be controlled? As moreover the approach to be used will be quite different from what is needed in the overdispersion case, we prefer to treat risk-adjusted negative binomial charts in a separate, forthcoming paper.

In Sect. 2 we demonstrate how the extension to the overdispersion case can be made from the negative binomial chart. Next, Sect. 3 is devoted to the performance of the new chart during OoC. In Sect. 4 the estimation aspects are covered. Finally, the procedure is summarized in Sect. 5.

2 Overdispersion

In the homogeneous case we have that $D_1, D_2, \ldots$ is a sequence of i.i.d. r.v.'s, with $Pr(D_1 = 1) = 1 - Pr(D_1 = 0) = p$ during IC. Once the process goes OoC, the failure probability p is replaced by θp for some $\theta > 1$ and a signal should follow as soon as possible. (Note that $\theta > 1$ is of primary interest, but a two-sided version can be derived in a completely similar way.). The 'time-between-events' approach means that we do not work with fixed-length blocks of D's, but instead wait each time till the rth failure occurs, for some $r \geq 1$. Let $X_i, i = 1, 2, \ldots$ be the successive numbers of D's involved, then these X_i clearly are i.i.d. copies of a negative binomial r.v. $X_{r,p}$ such that

$$Pr(X_{r,p} = k) = \binom{k-1}{r-1} p^r (1-p)^{k-r}, \tag{2.1}$$

where $k = r, r+1, \ldots$. Unless confusion might occur, we suppress the indices whenever possible, here as well as in the sequel, and thus simply write X instead of $X_{r,p}$. As $\theta > 1$, a signal should result when an rth failure arrives too soon, i.e. at the first time an $X_i \leq n$, for some suitable lower limit $n = n_{r,p}$.

Next we drop the assumption of homogeneity, according to which the D_i were identically distributed. Instead, each D_i has its own p_i, but we have no further knowledge about the underlying mechanism. All that is clear is that overdispersion causes an inadequate fit for the single parameter homogeneous model. Hence we need to consider a larger parametric family, by at least adding one (overdispersion) parameter. Clearly, this wider family cannot be expected to be 'true' either: it also remains an approximation of the underlying unknown structure. But, being wider, it should provide a better approximation.

Bearing the above in mind, we proceed as follows. In the homogeneous case, stopping at the rth failure led to the negative binomial $X_{r,p}$ from (2.1). To incorporate overdispersion, let P be a r.v. on $(0, 1]$ (or, more generally, on $(0, \infty)$, with $Pr(P > 1)$ negligible) such that

$$E\frac{p}{P} = 1, \ \mathrm{var}\frac{p}{P} = \tau, \tag{2.2}$$

where p is interpreted as the average failure rate and $\tau \geq 0$ is the overdispersion parameter. Typically, τ will not be really large, but also not sufficiently small to be negligible. Nevertheless, homogeneity (i.e. $\tau = 0$) is included as a boundary case. As our stopping time we now use $X_{r,P}$, i.e., given $P = p^*$, it is distributed as X_{r,p^*}. A straightforward calculation shows that

$$EX_{r,P} = \frac{r}{p}, \mathrm{var}(X_{r,P}) = \mathrm{var}(X_{r,p}) + \frac{r(r+1)\tau}{p^2} = \frac{r}{p^2}\{1 - p + (r+1)\tau\}. \tag{2.3}$$

Hence the relative increase due to overdispersion is $\approx (r+1)\tau$, expressing the joint effect of the length of the waiting sequence and the variation in failure rates.

Next we use these new r.v.'s to extend the basic homogeneous model as follows: once again we consider a sequence of i.i.d. r.v.'s, but now these will be copies of $X_{r,P}$ rather than of $X_{r,p}$. In other words, for each 'time-between-events'-sequence of length r, a new realization of P is chosen independently. As already argued above in general terms, this is just a modeling step, without the intention of precisely grasping the true underlying structure. Acting as if the basic sequence of D_i's conveniently selects a new value of P exactly if and only if an rth failure occurs, clearly is a simplification of reality. The point is that it is a considerably less stringent simplification than assuming homogeneity.

The obvious advantage of the parameterization above is that it allows us to keep using the results from the negative binomial case. In particular, given $P = p^*$, we have $Pr(X_{r,p^*} \leq n) \approx Pr(Z_{np^*} \geq r)$, where Z_{np^*} is a Poisson r.v. with parameter np^*. Taking expectations w.r.t. P thus leads to

$$Pr(X_{r,P} \leq n) \approx Pr(Z_{nP} \geq r). \tag{2.4}$$

Note that with (2.4) we have arrived at a classical overdispersion setup: a Poisson r.v. Z with random parameter. For the next modeling step, by far the most prominent choice [see e.g. Poortema (1999)] is to let this random parameter be Gamma distributed, resulting in a (shifted) negative binomial r.v.. To be more precise, let $G(\zeta, \eta)$ denote the gamma distribution with density

$$f_G(x) = \eta^{\zeta} x^{\zeta-1} \frac{e^{-\zeta x}}{\Gamma(\zeta)}, \quad x > 0, \tag{2.5}$$

and let Y_{n^*,p^*} denote a binomial r.v. with parameters n^* and p^*. Then we have for the present setup:

Lemma 2.1 *Let $\tau > 0$. If P is $G(2+\tau^{-1}, (1+\tau^{-1})/p)$, then (2.2) holds. Moreover, for this choice (2.4) specializes to*

$$Pr(X_{r,P} \leq n) \approx Pr\left(Y_{(r+1+\tau^{-1}),1/\{1+(1+\tau^{-1})/(np)\}} \geq r\right). \tag{2.6}$$

Proof See the Appendix.

Remark 2.1 For $\tau \to 0$, clearly $n^* = r + 1 + \tau^{-1} \to \infty$ and $p^* = 1/\{1 + (1 + \tau^{-1})/(np)\} \to 0$, while $n^* p^* \to np$. Hence in this case the binomial approximation from (2.6) boils down to the Poisson approximation $Pr(Z_{np} \geq r)$ used in the homogeneous case. The difference between these two approximations thus precisely reflects the overdispersion effect we want to quantify. □

Using Lemma 2.1, we can now obtain an approximation to the lower limit n. As a criterion we shall use that, for some small $\alpha > 0$,

$$Pr(X_{r,P} \leq n) = r\alpha. \tag{2.7}$$

In this way, the average run length (ARL) during IC will attain the same value $r/(r\alpha) = 1/\alpha$ for all r, thus allowing a fair comparison among the negative binomial charts for $r \geq 1$. Solving n numerically from $Pr\left(Y_{(r+1+\tau^{-1}),1/\{(1+(1+\tau^{-1})/(np)\}} \geq r\right) = r\alpha$ for given p, r, α and τ is not difficult, but a further approximation step is most convenient. We have:

Lemma 2.2 *The lower limit $n = n_\tau$ such that $Pr\left(Y_{(r+1+1/\tau),1/\{1+(1+1/\tau/(np)\}} \geq r\right) = r\alpha$ can be approximated for $p \leq 0.01$, $r \leq 5$, $\alpha \leq 0.01$ and $0 < \tau \leq 1/(r+1)$ by*

$$\tilde{n}_\tau = \alpha_{r\tau} \frac{1 + \zeta_{r\tau}}{p}, \text{ with } \alpha_{r\tau} = v\left(\frac{r\alpha}{\binom{v+r}{r}}\right)^{\frac{l}{r}} \tag{2.8}$$

and moreover $\zeta_{r\tau} = \alpha_{r\tau}(v+r+1)/\{v(r+1)\} + \frac{1}{2}(\alpha_{r\tau}^2[(3r+5)(v+r+1)^2/\{(r+1)^2(r+2)v^2\} - (v+r+1)/\{(r+2)v^2\}]$, with $v = 1 + \tau^{-1}$.

Proof See the Appendix.

Remark 2.2 As concerns the behavior for various τ, we note:

(i) *(upper bound)*. For $\tau = 1/(r+1)$ the variance has already doubled [cf. (2.3)]. Beyond this level simply adapting the homogeneous approach probably no longer suffices and more detailed information should be gathered (cf. the remarks about risk adjusted charts from the Introduction).

(ii) *(lower bound)*. For $\tau \to 0$, the quantities from (2.8) converge to their counterparts from Lemma 2.1 in Albers (2010) (cf. Remark 2.1). Hence for really small τ the additional effort to accommodate overdispersion is no longer worthwhile and a lower bound like $\tau \geq 0.05/(r+1)$ might be added as well. We did not do this because technically it is not needed, but in the estimation part (Sect. 4) it will reoccur.

(iii) *(given τ)*. Using some algebra, it is not difficult to verify in (2.8) that $\alpha_{r\tau}$ decreases in τ, and that this is also true for $\alpha_{r\tau}^2(v+r+1)/\{v(r+1)\}$, the next coefficient in the expansion for $\tilde{n}_\tau$. Hence $\tilde{n}_\tau$ typically decreases in τ, as should be the case, because overdispersion has a widening effect and thus forces us to lower the control limit $\tilde{n}_\tau$ in comparison to the result $\tilde{n}_0$ from the homogeneous case. □

Next we check the quality of $\tilde{n}_\tau$ as an approximation for n_τ. (See the Appendix for some details on how n_τ can be computed). In Table 1 below some illustrative values are collected. Following Albers (2010), we use $\alpha = 0.001$, 0.005 and 0.01. For the present purpose, the focus no longer is on comparing the various values of r, so we can restrict ourselves to representative values like $r = 3$ and $r = 5$. The emphasis now is on the relative overdispersion increase $(r+1)\tau$, for which we let the values range from 0 (homogeneous case) to 1 (upper limit in Lemma 2.2). The value of p (as long as it is at most 0.01) fortunately has no impact on the approximation quality, which allows us to summarize the results in terms of $n_\tau p$ and $\tilde{n}_\tau p$.

Table 1 Comparison of the approximation $\tilde{n}_\tau p$ from (2.8) to the exact $n_\tau p$ (cf. A.3), for various α, r and $(r+1)\tau$

$\alpha\backslash\beta$	0		0.05		0.1		0.2		0.5		1	
$r = 3$												
0.001	0.282	0.281	0.275	0.275	0.269	0.269	0.258	0.258	0.234	0.234	0.206	0.206
0.005	0.509	0.506	0.497	0.496	0.487	0.486	0.469	0.467	0.427	0.425	0.380	0.378
0.01	0.665	0.660	0.652	0.647	0.639	0.634	0.616	0.611	0.562	0.557	0.503	0.497
$r = 5$												
0.001	1.08	1.07	1.06	1.05	1.04	1.03	1.00	0.99	0.91	0.90	0.81	0.80
0.005	1.62	1.58	1.59	1.55	1.57	1.52	1.52	1.47	1.40	0.135	1.25	120
0.01	1.97	1.88	1.94	1.86	1.91	1.82	1.85	1.77	1.71	1.62	1.55	1.45

The first value is $n_\tau p$; the second one is $\tilde{n}_\tau p$

In Albers (2010) it was concluded that the approximation performs quite well over the region considered. Fortunately, Table 1 shows that this conclusion can be extended to the case of positive $(r+1)\tau$, all the way to the upper limit 1. Another important result from Table 1 is the observation that the effect of $(r+1)\tau$ indeed can be considerable: as it grows, the resulting values decrease quite a bit in comparison to those for the homogeneous case $\tau = 0$. Remember that this decrease serves to accommodate the overdispersion effect and to maintain the value of FAR during IC at $r\alpha$. By way of illustration we show in Table 2 what happens to this FAR if the overdispersion is ignored and $n_\tau p$ for $\tau = 0$ is used while in fact τ is positive. Indeed, the realized FAR can be doubled, or even tripled, if overdispersion effects become substantial, thus producing on the average far too short runs during IC.

To illustrate that application of the resulting chart is still quite simple, we conclude this section with:

Example 2.1 Suppose an ARL of 200 is considered acceptable, i.e. $\alpha = 0.005$ is chosen. If we want to decide about stopping or continuing at each third failure, we should use $r = 3$. In the homogeneous case [cf. Example 2.1 from Albers (2010)], we used n such that $Pr(Z_{np} \geq 3) = 0.015$ here, leading to $np = 0.509$ (or $\tilde{n}p = 0.506$). However, assume now that in fact $\tau = 1/8$, and thus $(r+1)\tau = 1/2$. According to Table 2, using the homogeneous np would produce $FAR = 0.0234$ rather than 0.0150. Hence we proceed by noting that $1+\tau^{-1} = 9$, and thus obtain $n_\tau p$ from solving $Pr(Y_{12,1/(1+9)/(np))} \geq 3) = 0.015$ [cf. (A.3)] or, more directly, $\tilde{n}_\tau p$ from (2.8), leading to $n_\tau p = 0.427$ and $\tilde{n}_\tau p = 0.425$ (cf. Table 1). To complete the example, fix a value of p as well, e.g. by letting $p = 0.001$. During IC, the third failure should then on average arrive after 3,000 observations. In the homogeneous case, action is taken if this already happens before at most 509 (or 506) observations. Taking the overdispersion into account now actually lowers these limits to 427 (or 425) in the present case. □

3 The OoC situation

In this section we let the process go OoC, in the sense that p is replaced by θp, for some $\theta > 1$. Hence $ARL = r/Pr(X_{r,\theta P} \leq n) \approx r/Pr(Z_{\theta nP} \geq r)$ [cf. (2.4)]. If a r.v. T is $G(\zeta, \eta)$, then θT is $G(\zeta, \eta/\theta)$, and in analogy to (2.6) we thus obtain that $Pr(X_{r,\theta P} \leq n) \approx Pr(Y_{(r+1+1/\tau),1/\{1+(1+\tau^{-1})/(\theta np)\}} \geq r)$. Consequently, under

Table 2 Realized FAR's (in %) when using the homogeneous values for $n_0 p$ for various α, r and $(r+1)\tau$

	$r=3$						$r=5$					
$\alpha \backslash \beta$	0	0.05	0.1	0.2	0.5	1	0	0.05	0.1	0.2	0.5	1
0.001	0.300	0.322	0.341	0.382	0.501	0.693	0.500	0.546	0.590	0.681	0.973	1.49
0.005	1.50	1.59	1.68	1.85	2.34	3.07	1.50	2.68	2.85	3.20	4.21	5.83
0.01	3.00	3.16	3.32	3.62	4.50	5.75	5.00	5.30	5.58	6.14	7.76	10.1

overdispersion we arrive at

$$ARL = ARL_{r,\theta} = \frac{r}{PR(Y_{(r+1+\tau^{-1}),1/\{1+(1+\tau^{-1})/(\theta np)\}} \geq r)} \tag{3.1}$$

with n such that $Pr(Y_{(r+1+1/\tau),1/\{1+(1+\tau^{-1})/(np)\}} \geq r) = r\alpha$. Hence, just as in the homogeneous case, going out of control leads to replacement of the relevant np by θnp. Not surprisingly, this means that the corresponding result from Albers (2010) can be adapted in a straightforward manner to

Lemma 3.1 *The ARL from (3.1) can be approximated for* $p \leq 0.01$, $r \leq 5$, $\alpha \leq 0.01$, $(r+1)\tau \leq 1$ *and* $3/2 \leq \theta \leq 4$ *by* $A\tilde{R}L = A\tilde{R}L_{r,\theta,\tau} =$

$$\frac{r}{1 - \frac{v}{(v+\theta\alpha_{r\tau})^{v+r}}\left[1 + \frac{\theta\alpha_{r\tau}(v+r)}{v} + \cdots + \binom{v+r}{r-2}(\frac{\theta\alpha_{r\tau}}{v})^{r-2} + \binom{v+r}{r-1}(\frac{\theta\alpha_{r\tau}}{v})^{r-1}\{1 - \frac{\theta\alpha_{r\tau}\xi_{r\tau}(v+1)}{v+\theta\alpha_{r\tau}[1+\xi_{r\tau}]}\}\right]}, \tag{3.2}$$

with $\alpha_{r\tau}$, $\zeta_{r\tau}$ *and* v *as in* (2.8).

Proof Apply the method of Lemma 3.1 from Albers (2010) (which is the boundary case of (3.2) as $\tau \to 0$] to the relevant binomial rather than Poisson probabilities. □

The range of values of interest for θ obviously remains the same as in Albers (2010). Just as in that paper, we are interested in the quality of the approximation provided, but now the focus is on the behavior with respect to τ. In Table 3 some illustrative values are collected, with α and r as in Tables 1 and 2 and θ as in Table 3 from

Table 3 Comparison of $A\tilde{R}L$ from (3.2) to ARL from (3.1) for various α, r, τ and θ

$\alpha\backslash\theta$	3/2		2		3		4	
$r = 3$								
0.001	329	338	154	162	55.7	61.3	28.7	32.7
	332	344	155	164	56.2	62.1	28.9	33.0
0.005	71.2	74.5	36.0	39.1	15.1	17.5	9.04	10.7
	73.4	77.9	36.9	40.6	15.4	17.9	9.10	10.9
0.01	37.6	39.7	20.0	22.0	9.32	10.9	6.04	7.27
	39.3	42.2	20.7	23.3	9.47	11.2	6.06	7.37
$r = 5$								
0.001	203	224	73.7	88.0	22.2	29.1	11.6	15.7
	233	160	82.1	69.6	23.5	25.4	11.8	14.4
0.005	49.8	56.3	21.9	26.8	9.31	12.1	6.44	8.22
	61.7	59.0	25.4	28.6	9.71	12.7	6.31	8.42
0.01	28.2	32.1	13.9	17.0	7.12	8.96	5.60	6.74
	36.3	39.4	16.2	20.2	7.21	9.87	5.30	7.05

In each 2×2 cell the upper values are ARL's and the lower ones $A\tilde{R}L$'s, while the left column is for $\tau = 0$ (homogeneity) and the right one for $(r+1)\tau = 1$

Albers (2010). Since the behavior in τ is again monotone (cf. Tables 1 and 2), we just present the boundary cases $\tau = 0$ and $(r+1)\tau = 1$.

Several interesting observations can be made from Table 3. As expected, the required numbers of observations increase as $(r+1)\tau$ goes from 0 to 1. Do note that this fact should not be interpreted as a 'drawback' of the adjusted charts, in the sense that avoiding this adjustment would in fact have produced a lower ARL and thus a better OoC performance. From Table 2 it is evident that such an 'improvement' can only be obtained by cheating on the requirement that $ARL = 1/\alpha$ during IC. Nevertheless, it is gratifying to observe as well that the impact of changing τ is much smaller under OoC than under IC. In the latter case, Table 2 shows that even tripling of the intended value can occur, while the relative increase in Table 3 is considerably smaller. Note that this phenomenon is of a general nature and by no means special for the present situation. In addition, Table 3 shows that in general the approximation works well in the region considered, with again a decreasing quality as $r\alpha$ increases. Moreover, observe that for small α and θ at $r = 5$ the approximation no longer increases as $(r+1)\tau$ goes from 0 to 1, which also indicates that here the limits of its usefulness are reached.

Yet another conclusion is that the pattern with respect to the optimal choice of r for given θ obviously hardly changes in going from the homogeneous case $\tau = 0$ to the opposite end at $(r+1)\tau = 1$. Consequently, there is no need to adapt the analysis from Albers (2010) at this point, and we can stick to the rule of thumb from that paper: for given α and θ the value r^{opt} that minimizes ARL_r is adequately approximated by

$$\min(5, \tilde{r}^{\text{opt}}), \tag{3.3}$$

where $\tilde{r}^{\text{opt}} = 1/\{\alpha(2.6\theta + 2) + 0.01(4\theta - 3)\}$ [cf. Table 3.2 from Albers (2010)]. The reason for the truncation of $\tilde{r}^{\text{opt}}$ suggested in (3.3) is twofold: (i) the main part of the improvement over the geometric chart usually is already achieved within the range $2 \le r \le 5$; (ii) having to collect a really large number of failures before being allowed to stop, might be considered undesirable in practice. To illustrate matters, we conclude the present section with:

Example 3.1 Using Example 2.1 as a starting point, let once more $\alpha = 0.005$, $p = 0.001$ and $r = 3$. Homogeneity in this situation gave $np = 0.509$ (or 0.506) and thus $n = 509$ (or 506). Suppose now that in fact $\tau = 1/4$, i.e. $(r+1)\tau = 1$, then during IC this choice would actually produce $FAR = 3.07\%$, instead of 1.50%. Hence the corresponding ARL would be less than 100, instead of the intended 200. Consequently, we definitely prefer to repair this defect by lowering our limit to $n = 380$ (or 378). The price for this correction during OoC boils down at $\theta = 4$ to an increase in ARL from 9.04 to 10.7 (or from 9.10 to 10.9), which seems quite moderate. Even after correction, 3 to 4 blocks of 3 failures on the average will suffice for a signal to occur.

Next observe that (3.3) suggests $r = 5$ as optimal choice for $\alpha = 0.005$ and $\theta = 4$. Then the lower limit $n = 1,620$ (or 1,580) should be lowered to $n = 1,280$ (or 1,200), in order to avoid a rise of the $IC - FAR$ from 2.50 to 5.83%. As a consequence, the $OoC - ARL$ at $\theta = 4$ will rise from 6.44 to 8.22 (or from 6.31 to 8.42).

Indeed some further improvement over $r = 3$ is achieved: 1 to 2 blocks of 5 failures will now suffice on average.

Finally, to illustrate that most of the gain with respect to the geometric chart (i.e. $r = 1$) typically is achieved within the range $2 \leq r \leq 5$, note the following. The geometric chart has $ARL \approx 1/(\theta\alpha)$ [see (2.2) in Albers (2010)], which means an ARL of about 50 here. The step towards $r = 3$ gives the main reduction to 9.04, with a slight further improvement for $r = 5$ to 6.44. The latter two values are those for the homogeneous case. Accommodating overdispersion means a renewed increase to 10.7 and 8.22, respectively, which is very mild compared to the starting value of 50. Hence also in this respect, the price for correcting for overdispersion seems quite fair. □

4 The estimated chart

Typically the underlying parameters of the chart will be unknown in practice. In the present setup not only the failure rate p is involved, but also the overdispersion parameter τ from (2.2). Hence these will have to be estimated and a Phase I sample is needed before monitoring can start. Let m be the size of such a sample, in the sense that we observe the sequence $D_1, D_2, \ldots$ until m failures have been gathered. Note that m does not depend on the r we choose: in this way, also with respect to estimation, fairness in comparing charts for different r is preserved. Also observe that the r.v.'s involved are typically not simply distributed as $X_{r,p}$ from (2.1) for the homogeneous case, but also not necessarily as $X_{r,P}$ from (2.3), since this latter choice was proposed as a convenient modeling step (cf. the discussion in Sect. 2). Hence we prefer to adopt the following general notation: for simplicity (and without essential loss of generality), let $k = m/r$ be an integer, then our Phase I sample consists of k r.v.'s $Y_{r,p}$. Here each $Y_{r,p}$ is an overdispersed waiting time till the rth failure, so let us use here as well [cf. (2.3)] the notation

$$EY_{r,p} = \frac{r}{p}, \quad \mathrm{var}(Y_{r,p}) = \frac{r}{p^2}\{1 - p + (r+1)\tau\}. \tag{4.1}$$

In this way, for both $Y_{r,p}$ and $X_{r,P}$, the relative increase due to overdispersion is denoted by $(r+1)\tau/(1-p) \approx (r+1)\tau$.

For briefness' sake denote the k $Y_{r,p}$'s from Phase I by $Y_1, \ldots, Y_k$ and let

$$Y^* = m^{-1}\Sigma_{i=1}^{k} Y_i, \; S_r^2 = (m-r)^{-1}\Sigma_{i=1}^{k}(Y_i - rY^*)^2, \tag{4.2}$$

then we suggest the following estimators (see the Appendix for some details)

$$\hat{p} = \frac{1}{Y^*}, \; \hat{\tau} = \max\left(0, \frac{S_r^2}{(Y^*)^2} - 1\right) \Big/ (r+1). \tag{4.3}$$

The maximum in (4.3) has been included since nonpositive values of $S_r^2/(Y^*)^2 - 1$ can occur. However, this is a negligible complication, because it will typically only happen

if the underlying τ is really small. Such τ are not at all interesting and taking the trouble to accommodate the overdispersion effect can be reserved for e.g. $\tau \geq 0.05/(r+1)$ [cf. Remark 2.2 (ii)]. Hence the proper reaction in practice to finding such a nonpositive value is to refrain from additional effort, i.e. to stick to the homogeneous approach. That is precisely what (4.3) does: $\hat{\tau} = 0$ in that case.

Basically, the above is all that is needed to transform the chart into its estimated version: just replace p and τ in Sects. 2 and 3 by their estimated counterparts $\hat{p}$ and $\hat{\tau}$. For example, instead of the lower limit $n = n_\tau$ solving $Pr(Y_{(r+1+\tau^{-1}),1/\{1+(1+\tau^{-1})/(np)\}} \geq r) = r\alpha$, we now have $n = \hat{n}_\tau$ such that

$$Pr\left(Y_{(r+1+\hat{\tau}^{-1}),1/\{1+(1+\hat{\tau}^{-1})/(n\hat{p})\}} \geq r\right) = r\alpha. \tag{4.4}$$

Likewise, $\tilde{n}_\tau$ from (2.8) becomes $\hat{\tilde{n}}\tau = \alpha_{r\hat{\tau}}(1+\zeta_{r\hat{\tau}})/\hat{p}$ (just substitute $\hat{v} = 1+\hat{\tau}^{-1}$ for $v = 1+\tau^{-1}$ everywhere in $\alpha_{r\tau}$ and $\zeta_{r\tau}$). Once such an estimated lower limit $\hat{n}_\tau$ (or $\hat{\tilde{n}}_\tau$) has been obtained from the Phase I sample, the actual monitoring can start: each time we wait till the rth failure, and if this occurs at or before this lower limit, a signal is given. Hence, straightforward application of the estimated chart remains easy.

However, it remains to note that as a consequence of the estimation step the performance characteristics FAR and ARL will now be stochastic, rather than fixed at $r\alpha$ and $1/\alpha$, respectively. To be able to control this effect, we include the possibility to apply a small correction c to the estimated limit $\hat{n}_\tau$ from (4.4):

$$\hat{n}_{\tau,c} = \hat{n}_\tau(1-c) \quad \text{and} \quad \widehat{FAR}_c = Pr(X_{r,P} \leq \hat{n}_{\tau,c} | Y^*, S_r^2). \tag{4.5}$$

Hence for $c = 0$ we again have the uncorrected case: $\hat{n}_{\tau,0} = \hat{n}_\tau$. A quantity of interest now e.g. is the exceedance probability $Pr(\widehat{FAR}_0 > r\alpha(1+\varepsilon))$ for the uncorrected case, and moreover the value of c such that, for some prescribed small δ

$$Pr(\widehat{FAR}_c > r\alpha(1+\varepsilon)) \leq \delta. \tag{4.6}$$

In this connection note that $Pr(\widehat{ARL}_c < (1-\varepsilon)/\alpha) = Pr(r/\widehat{FAR}_c < (1-\varepsilon)/\alpha)) = Pr(\widehat{FAR}_c > r\alpha(1+\tilde{\varepsilon}))$, where $\tilde{\varepsilon} = \varepsilon/(1+\varepsilon)$. Hence control of $\widehat{FAR}_c$ through (4.6) automatically provides that of $\widehat{ARL}_c$, and vice versa.

To evaluate this exceedance probability, as well as to find c such that (4.6) holds, we introduce

$$U = \frac{p}{\hat{p}} - 1, \; W = -\frac{(r-np)(\hat{\tau}-\tau)}{\{1+(r+1)\tau\}(1+\tau)}, \tag{4.7}$$

and denote the standard deviation of $(U+W)$ by $\sigma_{(U+W)}$. Moreover, let

$$\gamma_\tau = \frac{Pr(Y_{(r+1+1/\tau),1/\{1+(1+1/\tau)/(np)\}} = r)}{r\alpha\{1+np/(1+\tau^{-1})\}}, \tag{4.8}$$

and write u_δ for the upper δ-point of the standard normal d.f. Φ, i.e. $1 - \Phi(u_\delta) = \delta$. Then we have

Lemma 4.1 *The uncorrected exceedance probability satisfies*

$$P(\widehat{FAR}_0 > r\alpha(1+\varepsilon)) \approx 1 - \Phi\left(\frac{\varepsilon}{\gamma_\tau r \sigma_{(U+W)}}\right), \tag{4.9}$$

while equality in (4.6) *is achieved by using* $\hat{n}_{\tau,c}$ *from* (4.5) *with*

$$c = \sigma_{(U+W)} u_\delta - \frac{\varepsilon}{\gamma_\tau r}. \tag{4.10}$$

Moreover γ_τ *satisfies* $1 - (r + 2 + \tau^{-1})/\{(1 + (1 + \tau^{-1})/(np))(r + 1)\} < \gamma_\tau < 1/\{1 + np/(1 + \tau^{-1})\}$.

Proof See The Appendix.

Once again, letting $\tau \to 0$ reproduces the results from the homogeneous case. In particular, $\gamma_\tau \to \gamma$ with $1 - \lambda/(r+1) < \gamma < 1$ and $\sigma_{(U+W)} \to \sigma_U$, which for $\tau = 0$ simply equals $m^{-1/2}$ to first order. In the present case, some effort is needed to obtain $\sigma_{(U+W)}$. The expressions involved are more complicated and an additional estimation step is required. For some details, see the Appendix. In addition to $\hat{p}$ and $\hat{\tau}$, moment estimators $\hat{\mu}_j = k^{-1}\Sigma_{i=1}^k (Y_i - \overline{Y})^j$, $j = 3$ and 4, are needed. The resulting $\hat{\sigma}_{(U+W)}$ still is of order $m^{-1/2}$, implying that the correction c from (4.10) will indeed be small if the Phase I sample size m is sufficiently large.

5 Summary

For convenience, we summarize the application of the overdispersion chart as discussed in the previous sections:

1. Select a desired $IC - ARL = 1/\alpha$ and a degree of change $\theta > 1$ for p during OoC that should be optimally protected against.
2. Apply rule of thumb (3.3) to obtain the best r for this α and θ.
3. For known p and τ, compute the lower limit $\tilde{n}_\tau$ from (2.8).
4. If desired, use (3.2) to check whether the $OoC - ARL$ is satisfactory.
5. Start monitoring: at each rth failure, signal if $\leq \tilde{n}_\tau$ observations were observed.
6. If p and τ are unknown, select an $m = kr$ and first collect $Y_1, \ldots, Y_k$.
7. Apply (4.2) and (4.3) to obtain $\hat{p}$ and $\hat{\tau}$. Use these values at step 3.
8. If desired, analyze exceedance probabilities through (4.9) and (4.10).

As argued in the Introduction (also see the references mentioned there), this chart is especially suited for medical applications, where failures are supposed to be quite rare (congenital malformations, cancer incidence, surgical errors) and overdispersion

is quite common [varying patient characteristics, multiple units (hospitals, general practices)]. Woodall (2006) states that many issues on how to best adjust control charts for overdispersion remain unresolved. The present proposal is meant to offer a contribution in this respect. In particular, it is easy to implement, uses the best r for chosen θ, allows a simple check of the resulting $OoC - ARL$, as well as an appraisal of the effect if overdispersion had been ignored. Moreover, the—typically needed—estimated version is simple as well, while appraisal of and/or correction for the estimation effects are available.

The properties just mentioned have been illustrated already in Examples 2.1 and 3.1. Hence to avoid repetition, here we just give a short numerical example of the summary above. Suppose some ongoing stream of health care or public health surveillance data are gathered in the form of waiting times till observed failures. An $IC - ARL = 200$ is chosen and the focus is on possible quadrupling of the failure rate, i.e. $\theta = 4$ (step 1). This results in $r = 5$: waiting till each fifth failure before deciding on a signal (step 2). But as p and τ are unknown, first a Phase I sample consisting of $m = 150$ failures is collected, leading to $k = 30$ r.v.'s Y_i (step 6). Suppose these produce $\hat{p} = 0.002$ and $\hat{\tau} = 1/12$, indicating that overdispersion cannot be neglected (step 7). Using these values, compute $\hat{\tilde{n}}_\tau \hat{p} = 1.35$ and thus $\hat{n}_\tau = 675$ (step 3). In principle, monitoring starts right now: a signal arises as soon as a fifth failure arrives at or before the 675th patient. (If desired, first check $OoC - ARL$ (step 4) and/or exceedance probabilities (step 8). Modify the choice of α, θ and/or m if considered necessary, and repeat the steps involved).

Appendix

Proof of Lemma 2.1 Let T be a r.v. with d.f. $G(\zeta, \eta)$ (cf. (2.5)), then $E(1/T) = \eta/(\zeta-1)$, $E(1/T)^2 = \eta^2/\{(\zeta-1)(\zeta-2)\}$, and thus $\text{var}(1/T) = \{\eta/(\zeta-1)\}^2/(\zeta-2)$. For the special case $T = P$, we have $\zeta = 2 + \tau^{-1}$ and $\eta = (1 + \tau^{-1})/p$, and thus $E(1/P) = 1/p$, $\text{var}(1/P) = \tau/p^2$. Hence (2.2) indeed holds. From (2.5) it is immediate that $Pr(Z_T = k) = \int_0^\infty Pr(Z_X = k) f_G(x)dx = \Gamma(\zeta + k)/\{k!\Gamma(\zeta)\}\{1/(\eta + 1)\}^k\{\eta/(\eta + 1)\}^\zeta$, $k = 0, 1, \ldots$. This means that $Z_T + \zeta$ is distributed as the negative binomial $X_{\zeta,\eta/(\eta+1)}$ from (2.1). Consequently,

$$\begin{aligned} Pr(Z_T \geq r) &= Pr(X_{\zeta,\eta/(\eta+1)} > \zeta + r - 1) = Pr(Y_{\zeta+r-1,\eta/(\eta+1)} < \zeta) \\ &= Pr(Y_{\zeta+r-1,1/(\eta+1)} \geq r). \end{aligned} \tag{A.1}$$

As $T = nP$ in (2.4), application of (A.1) with $\zeta = 2 + \tau^{-1}$ and $\eta = (1 + \tau^{-1})/(np))$ produces the desired result (2.6). □

Proof of Lemma 2.2 This is a straightforward extension of the proof of Lemma 2.1 from Albers (2010). There a result from Klar (2000) for Poisson probabilities is applied, which shows that the error committed by replacing $Pr(Z_{np} \geq r)$ by $\Sigma_{j=r}^{r+2} Pr(Z_{np} = j)$ is sufficiently small. But Klar (2000) contains a similar result for the binomial case, and this can be used here for $Y_{(r+1+\tau^{-1}),1/\{1+(1+\tau^{-1})/(np)\}}$ in precisely the same manner. The second step in that proof consists of expanding

$\Sigma_{j=r}^{r+2} Pr(Z_\lambda = j)$ w.r.t. $\lambda = np$ to third order. By equating the result obtained to $r\alpha$ and inverting w.r.t. λ, the desired expansion follows. Again, the same procedure, be it a bit more laborious, can be applied here. To provide some details, note that after the first step, we have the approximation

$$r\alpha = \binom{v+r}{r} \frac{\lambda^r v^v}{(v+\lambda)^{v+r}} \left\{ 1 + \frac{\lambda}{r+1} + \frac{(v-1)\lambda^2}{(r+1)(r+2)v} \right\}, \tag{A.2}$$

from which it is immediate that $(\alpha_{r\tau})^r = \lambda^r\{1 + O(\lambda)\}$, and thus $\lambda = \alpha_{r\tau}$ to first order. The refinement in (2.8) follows by solving (A.2) to third, rather than just first, order. □

Computation of $\mathbf{n}_\tau$. According to Lemma 2.2, we are looking for the solution $n = n_\tau$ of the equation

$$\begin{aligned} Pr(Y_{(r+1+\tau^{-1}),1/\{1+(1+\tau^{-1})/(np)\}} \geq r) &= Pr(X_{r,1/\{1+(1+\tau^{-1})/(np)\}} \leq r + 1 + \tau^{-1}) \\ &= r\alpha. \end{aligned} \tag{A.3}$$

By way of illustration, first consider the geometric case $r = 1$, where a direct approach is feasible. Here (A.3) boils down to $\alpha = Pr(X_{1,1/\{1+v/(np)\}} \leq v + 1) = 1 - \{1 - 1/(1+v/(np))\}^{v+1}$, with again $v = 1+\tau^{-1}$. Hence $(1+np/v)^{-(v+1)} = 1-\alpha$ and thus the solution $n_\tau p = v\{(1-\alpha)^{-1/(v+1)}\} - 1$ is readily obtained. Indeed, expanding this expression leads to $\tilde{n}_\tau p = v\alpha/(v+1)\{1+\frac{1}{2}(v+2)\alpha/(v+1)+(v+2)(2v+3)\alpha^2/[6(v+1)^2]\}$, which agrees with (2.8) for $r = 1$. In passing also observe the following. In the geometric case $r = 1$ we directly have that $Pr(X_{1,P} \leq n) = 1 - E(1-P)^n = 1-\Sigma_{k=0}^{n}\binom{n}{k}(-1)^k EP^k$, with $EP^k = p^k\{(v+1)\ldots(v+k)\}/v^k$, as P is $G(v+1, v/p)$-distributed (cf. Lemma 2.1). Using the Poisson approximation subsequently gives $Pr(X_{1,P} \leq n) \approx Pr(Z_{nP} \geq 1) = 1 - Ee^{-nP} = 1 - \{v/(v+np)\}^{v+1}$, which in its turn agrees with the result derived just above, using $Pr(X_{1,1/\{1+(1+\tau^{-1})/(np)\}} \leq v+1)$.

For $r > 1$, obtaining n_τ is less straightforward. Let $\xi = 1/(1 + v/(np))$, then for given ξ we have from (A.3) that $v + r = F_{r,\xi}^{-1}(r\alpha)$, the $r\alpha$th quantile of the negative binomial df $F_{r,\xi}$. (We shall use an interpolated version.) Consequently, we obtain

$$np = \left(\frac{\xi}{1-\xi}\right)\{F_{r,\xi}^{-1}(r\alpha) - r\}, \ \tau = \{F_{r,\xi}^{-1}(r\alpha) - r - 1\}^{-1}, \tag{A.4}$$

for given r, α and ξ. By adapting the value of ξ, selected values for $(r + 1)\tau$ can be obtained iteratively in (A.4), and thus the corresponding $n = n_\tau$ as well.

Proof of Lemma 4.1 The result for γ_τ follows by once more using Klar (2000). Together, (A.5) from Lemma A.1 below and (4.8) imply that $\widehat{FAR}_c \approx r\alpha(1 + \gamma_\tau r\{U + W - c\})$. Hence the exceedance probability from (4.6) to first order equals $Pr(\gamma_\tau r\{U + W - c\} > \varepsilon)$. As $U + W$ is asymptotically normal with mean 0 and variance $\sigma^2_{(U+W)}$, this probability approximately equals $1 - \Phi(\{c + \varepsilon/(\gamma_\tau r)\}/\sigma_{(U+W)})$. For $c = 0$, this produces (4.9). If instead the prescribed δ should result, $c + \varepsilon/(\gamma_\tau r)$ has to equal $\sigma_{(U+W)}u_\delta$, and hence c should be chosen as in (4.10). □

Lemma A.1 *it To first order* $\widehat{FAR}_c$ *equals*

$$r\alpha + \frac{r}{1 + np/(1+\tau^{-1})}\{U + W - c\}Pr(Y_{(r+1+\tau^{-1}),1/\{1+(1+\tau^{-1})/(np)\}} = r) \quad \text{(A.5)}$$

in which $n = n_\tau$ *solves* $Pr(Y_{(r+1+\tau^{-1}),1/\{1+(1+\tau^{-1})/(np)\}} \geq r) = r\alpha$.

Proof As $n = \hat{n}_\tau$ is such that $Pr(Y_{(r+1+\hat{\tau}^{-1},1/(1+\hat{\tau}^{-1})/(n\hat{p}))\}} \geq r)$ equals $r\alpha$ as well (cf. (4.4)), it follows that $(r + 1 + \hat{\tau}^{-1})/\{1 + (1 + \hat{\tau}^{-1})/(\hat{n}_\tau \hat{p})\}$ to first order equals $(r+1+\tau^{-1})/\{1+(1+\tau^{-1})/(n_\tau p)\}$. Consequently, $\hat{n}_\tau \hat{p}/(n_\tau p) - 1 \approx (r - n_\tau p)(\tau - \hat{\tau})/\{(1 + \tau)(1 + (r + 1)\hat{\tau}\}$, which agrees to first order with W in (4.7). Hence, in view of (4.5), $\hat{n}_{\tau,c}/n_\tau = \{\hat{n}_\tau \hat{p}/(n_\tau p)\}(p/\hat{p})(1 - c) \approx 1 + U + W - c$. In view of (A.3), this implies that $\widehat{FAR} \approx Pr(Y_{(r+1+\tau^{-1}),1/\{1+(1+\tau^{-1})/(np)\}} \geq r)$, where now $n = n_\tau(1 + U + W - c)$. Since $\partial Pr(Y_{n,p} \geq r)/\partial p = (p/r)Pr(Y_{n,p} = r)$, a first order expansion around $1/\{1 + (1 + \tau^{-1})/(n_\tau p)\}$ then produces the result in (A.5). □

Derivation of $\hat{\mathbf{p}}$ and $\hat{\tau}$. Clearly, $Y^* = r^{-1}\overline{Y}$, with $\overline{Y} = k^{-1}\Sigma_{i=1}^k Y_i$, and thus Y^* is just the average waiting time till the first failure, with $EY^* = 1/p$. Moreover, $S_r^2 = r^{-1}\tilde{S}_r^2$, where $\tilde{S}_r^2 = (k-1)^{-1}\Sigma_{i=1}^k (Y_i - \overline{Y})^2$, the sample variance of the Y_i's. Consequently, $E\tilde{S}_r^2 = \text{var}(Y_1)\{1 - [k(k-1)]^{-1}\Sigma\Sigma_{i\neq j}\rho(Y_i, Y_j)\}$. Obviously, if the Y_i are distributed as in (2.1) (i.e. homogeneity holds after all), all correlations involved will be 0. More important, however, is the fact that this remains true if the Y_i are distributed according to (2.3), i.e. as $X_{r,P}$. Then not only all underlying D's are independent, but also a new and independent P is drawn after each rth failure. Note that this observation indicates what will happen for general Y_i. Typically, the effect of the correlation terms in $E\tilde{S}_r^2$ will remain negligible, as the only contribution comes from carryover effects, due to carrying on for a while with the same p after an rth failure. Only if the stretches involved are too large, problems will arise in this respect. However, as stated before, under such circumstances a closer scrutiny of the underlying process seems indicated (risk adjustment methods etc.). The present approach focuses on the simple setup where the information available essentially consists only of waiting times till rth failures. Hence we may assume that $ES_r^2 \approx r^{-1}\text{var}(Y_1) \approx \{1 + (r + 1)\tau\}/p^2$ [cf. (4.2)]. Then it follows that $p = 1/EY^*$ and $\tau \approx \{ES_r^2/(EY^*)^2 - 1\}/(r + 1)$, leading to (4.3).

Estimation of $\sigma_{(U+W)}$. As before, we assume that possible dependencies between the Y_i are negligible. For their marginal distribution, we might use that of $X_{r,P}$ and accordingly express the 3rd and 4th central moments involved in terms of r, p and τ. However, the resulting expressions are rather complicated. Moreover, simplification by using expansion w.r.t τ only works quite locally, as the coefficients of the higher order terms tend to grow considerably. But, apart from these technical aspects, it seems better anyhow not to rely on such an assumption and to just use moment estimators like $\hat{\mu}_j = k^{-1}\Sigma_{i=1}^k (Y_i - \overline{Y})^j$ for μ_j, $j = 3$ or 4. Then we can proceed as follows: first note that W from (4.7) to first order can written as $-a\{(1 + U^*)/(1 + U)^2 - 1\}$,

where

$$a = \frac{r - np}{(r+1)\tau} \quad \text{and} \quad U^* = \frac{p^2 \tilde{S}_r^2}{r\{1 + (r+1)\tau\}} - 1. \tag{A.6}$$

Hence $U + W \approx (1 + 2a)U - aU^*$. From (4.7) and (4.3) it follows that $\sigma_U^2 = p^2 \text{var}(Y^*)$, which in view of (4.2) and (4.1) leads to $\sigma_U^2 = (p/r)^2 \text{var}(\overline{Y}) = (p/r)^2(r/p^2)\{1 - p + (r+1)\tau\}/k \approx m^{-1}(1 + (r+1)\tau)$. Consequently, σ_U^2 can be estimated by $m^{-1}(1 + (r+1)\hat{\tau})$. For $\text{Cov}(U, U^*)$ and $\sigma_{U^*}^2$ similar steps can be taken. We obtain that $\text{Cov}(U, U^*) \approx p^3/\{r^2(1 + (r+1)\tau)\}\text{Cov}(\overline{Y}, \tilde{S}_r^2) = p^3/\{r^2(1 + (r+1)\tau)\}\mu_3/k = m^{-1}p^3/\{r(1 + (r+1)\tau)\}\mu_3$ and $\sigma_{U^*}^2 \approx p^4/\{r^2(1 + (r+1)\tau)^2\}\text{Var}(\tilde{S}_r^2) = [p^4/\{r^2(1 + (r+1)\tau)^2\}\mu_4 - 1]/k = m^{-1}[p^4/\{r(1 + (r+1)\tau)^2\}\mu_4 - r]$. Hence $\sigma_{(U+W)}$ now readily follows, after which replacement of p, τ and μ_j by their respective estimators gives the desired $\hat{\sigma}_{(U+W)}$. Note that $\hat{\sigma}_{(U+W)}$ still is of order $m^{-1/2}$, implying that the correction c from (4.10) will indeed be small if the Phase I sample size m is sufficiently large.

References

Albers W (2010) The optimal choice of negative binomial charts for monitoring high-quality processes. J Stat Plan Inference 140:214–225

Bourke PD (1991) Detecting a shift in fraction nonconforming using run-length control charts with 100% inspection. J Qual Technol 23:225–238

Bourke PD (2006) The RL_2 chart versus the np chart for detecting upward shifts in fraction defective. J Appl Stat 33:1–15

Chen R (1978) A surveillance system for congenital malformations. J Am Stat Assoc 73:323–327

Chen R (1987) The relative efficiency of the sets and the $CUSUM$ techniques in monitoring the occurrence of a rare event. Stat Med 6:517–525

Christensen A, Melgaard M, Iwersen J, Thyregod P (2003) Environmental monitoring based on a Hierarchical Poisson-Gamma Model. J Qual Technol 35:275–285

Fang Y (2003) c-Charts, X-Charts, and the Katz family of distributions. J Qual Technol 35:104–114

Gallus G, Mandelli C, Marchi M, Radaelli G (1986) On surveillance methods for congenital malformations. Stat Med 5:565–571

Grigg O, Farewell V (2004a) An overview of risk-adjusted charts. J R Stat Soc A 167:523–539

Grigg O, Farewell V (2004b) A risk-adjusted sets method for monitoring adverse medical outcomes. Stat Med 23:1593–1602

Grigg OA, Spiegelhalter DJ, Jones HE (2009) Local and marginal control charts applied to methicillin resistant Staphylococcus aureus bacteraemia reports in UK acute Nat. Health Service trusts. J R Stat Soc A 172:49–66

Klar B (2000) Bounds on tail probabilities of discrete distributions. Probab Eng Infor Sci 14:161–171

Liu JY, Xie M, Goh TN, Ranjan P (2004) Time-between-events charts for on-line process monitoring. Int Eng Man Conf: 1061–1065

Marshall C, Best N, Bottle A, Aylin P (2004) Statistical issues in the prospective monitoring of health outcomes across multiple units. J R Stat Soc A 167:541–559

Ohta H, Kusukawa E, Rahim A (2001) A $CCC - r$ chart for high-yield processes. Qual Reliab Eng Int 17:439–446

Poortema K (1999) On modelling overdispersion of counts. Stat Neerl 53:5–20

Shaha SH (1995) Acuity systems and control charting. Qual Manage Health Care 3:22–30
Sonesson C, Bock D (2003) A review and discussion of prospective statistical surveillance in public health. J R Stat Soc A 166:5–21
Thor J, Lundberg J, Ask J, Olsson J, Carli C, Härenstam KP, Brommels M (2007) Application of statistical process control in healthcare improvement: systematic review. Qual Saf Health Care 16:387–399
Woodall WH (2006) The use of control charts in health care monitoring and public health surveillance. J Qual Technol 38:89–104
Wu Z, Zhang X, Yeo SH (2001) Design of the sum-of-conforming-run-length control charts. Eur J Oper Res 132:187–196
Xie M, Goh TN, Lu XS (1998) A comparative study of CCC and $CUSUM$ charts. Qual Reliab Eng Int 14:339–345
Yang Z, Xie M, Kuralmani V, Tsui K-L (2002) On the performance of geometric charts with estimated parameters. J Qual Technol 34:448–458
Zhang L, Govindaraju K, Bebbington M, Lai CD (2004) On the statistical design of geometric control charts. Qual Technol Quant Manage 1(2):233–243